Concentration inequalities for the exponential weighting method

Yu Golubev, D Ostrovski

Concentration Inequalities for the Exponential Weighting Method *

Golubev, Yu.† and Ostrovski, D. ‡

Abstract

The paper is concerned with recovering an unknown vector from noisy data with the help of a family of ordered smoothers [11]. The estimators withing this family are aggregated based on the exponential weighting method and the performance of the aggregated estimate is measured by the excess risk controlling deviation of the square losses from the oracle risk. Based on natural statistical properties of ordered smoothers, we propose a novel method for obtaining concentration inequalities for the exponential weighting method.

1 Introduction and main results

This paper deals with recovering an unknown vector $\mu \in \mathbb{R}^n$ from the noisy observations

$$Y_i = \mu_i + \sigma\xi_i, \quad i = 1, 2, \dots, n, \tag{1.1}$$

where ξ is a standard white Gaussian noise, i.e., ξ_i are i.i.d. Gaussian random variables with $\mathbf{E}\xi_i = 0$ and $\mathbf{E}\xi_i^2 = 1$. For the sake of simplicity it is assumed that the noise level $\sigma > 0$ is known. It is also assumed that n is large. Let us emphasize that all results in this paper can be extended to the case $n = \infty$ provided that $\mu \in \ell_2$.

Notice also that in spite of the obvious simplicity of this statistical model, it can cover a very wide class of statistical problems ranging from regression estimation to inverse problems (see, e.g., [11], [8]).

*This work is partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF Government grant, ag. 11.G34.31.0073; and RFBR research projects 13-01-12447 and 13-07-12111.

†Aix Marseille Université, CNRS, LATP, UMR 7353 and Institute for Information Transmission Problems, 39 rue F. Joliot Curie, 13453 Marseille, France.

‡Moscow Institute of Physics and Technology, Institutski per. 9, Dolgoprudny, 141700, Russsia.

In this paper, the quality of an estimate $\hat{\mu}(Y)$ is measured by the mean square risk

$$\rho(\hat{\mu}, \mu) = \mathbf{E}_\mu \|\hat{\mu}(Y) - \mu\|^2,$$

where $\mathbf{E}_\mu$ stands for the expectation with respect to the measure $\mathbf{P}_\mu$ generated by the observations (1.1) under given μ, and $\|\cdot\|$ is the standard Euclidean norm in $\mathbb{R}^n$.

In what follows it is assumed that we have at our disposal a family of linear estimates

$$\hat{\mu}^h_k(Y) = h_k Y_k, \ h \in \mathcal{H}, \tag{1.2}$$

where $\mathcal{H}$ is a given set of vectors in $\mathbb{R}^n$. This set consists of vectors with specific properties that will be described later on.

Our goal is to construct a good estimate of μ with the help of the family of linear estimates $\hat{\mu}^h(Y), \ h \in \mathcal{H}$. To explain what estimates might be viewed as good, let us notice that for given h, the risk of $\hat{\mu}^h(Y)$ is computed as follows :

$$R(\mu, h) \stackrel{\text{def}}{=} \mathbf{E}_\mu \|\hat{\mu}^h(Y) - \mu\|^2 = \|(1-h)\cdot\mu\|^2 + \sigma^2\|h\|^2, \tag{1.3}$$

where here and in what follows $\cdot$ means the component-wise multiplication of vectors in $\mathbb{R}^n$, i.e, $z = x \cdot y$ means that $z_k = x_k y_k, \ k = 1, \ldots, n$.

Of coarse, we want to find a method of combining estimates $\{\hat{\mu}^h(Y), \ h \in \mathcal{H}\}$ that provides an estimate with the minimal risk. If the family of estimates is reach enough, then we would be satisfied with estimators whose risks are close to the value

$$r(\mu, \mathcal{H}) = \min_{h\in\mathcal{H}} \Big\{ \|(1-h)\cdot\mu\|^2 + \sigma^2\|h\|^2 \Big\}.$$

This risk is often called oracle risk since it coincides with the risk of the following psudo-estimate

$$\hat{\mu}^*(Y) = h^*(\mu)\cdot Y, \qquad h^*(\mu) = \arg\min_{h\in\mathcal{H}} \Big\{ \|(1-h)\cdot\mu\|^2 + \sigma^2\|h\|^2 \Big\}. \tag{1.4}$$

Obviously, this object is not a real statistical estimate since it depends on the unknown vector μ.

A rather natural approach to constructing new estimates with the help of the available ones is to compute their convex combination

$$\bar{\mu}^w(Y) = \sum_{h\in\mathcal{H}} w_h(Y)\hat{\mu}^h(Y), \tag{1.5}$$

where $w_h(Y)$ are some nonnegative numbers called weights such that

$$\sum_{h\in\mathcal{H}} w_h(Y) = 1.$$

To simplify unessential technical details, here and below $\mathcal{H}$ is assumed to be finite.

The main issue in this method is evidently related to the selection of weights $w_h(Y)$. In this paper, we focus on the so-called exponential weights defined as follows :

$$w_h(Y) = \frac{\pi_h}{Z(Y)} \exp\left[-\frac{\bar{R}(Y,h)}{2\beta\sigma^2}\right], \tag{1.6}$$

where

$$Z(Y) \stackrel{\text{def}}{=} \sum_{h\in\mathcal{H}} \pi_h \exp\left[-\frac{\bar{R}(Y,h)}{2\beta\sigma^2}\right], \tag{1.7}$$

$$\bar{R}(Y,h) \stackrel{\text{def}}{=} \|\hat{\mu}^h(Y)\|^2 - 2\langle Y, \hat{\mu}^h(Y)\rangle + 2\sigma^2 \sum_{h\in\mathcal{H}} h_k, \tag{1.8}$$

is the empirical counterpart of $R(\mu,h)$ and β, π_h are positive numbers. It is clear that without loss of generality one may assume that $\sum_{h\in\mathcal{H}} \pi_h = 1$.

To explain heuristically how these weights can be obtained, let us simplify the initial statistical problem assuming that we have at our disposal the auxiliary observations

$$Y'_k = \mu_k + \sigma\xi'_k, \quad k = 1, 2\ldots, n, \tag{1.9}$$

where ξ' is independent of ξ in (1.1) a standard white Gaussian noise. These observations will be used to select weights $w_h(Y')$ in the estimate

$$\bar{\mu}^w(Y,Y') = \sum_{h\in\mathcal{H}} w_h(Y')\hat{\mu}^h(Y), \tag{1.10}$$

where $w_h(Y') \geq 0$ and $\sum_{h\in\mathcal{H}} w_h(Y') = 1$.

Apparently the approach to constructing new estimates with the help of the auxiliary sample were firstly studied and developed by A. Nemirovsky and, independently, by A. Catoni (see [15, 4]). Subsequently, these methods have been adapted for a wide class of statistical models (see, for example, [19], [16], [12], [17]).

Let us assume that Y is frozen. The idea in choosing $w_h(\cdot)$ in (1.10) is related to the so-called roughness penalty approach yielding the weights

$$\tilde{w}_h(Y') = \arg\min_{w_h}\left\{\frac{1}{2\epsilon^2}\left\|Y' - \sum_{h\in\mathcal{H}} w_h\hat{\mu}^h(Y)\right\|^2 + \beta K(w_h,\pi_h)\right\}, \qquad (1.11)$$

where

$$K(w_h,\pi_h) = \sum_{h\in\mathcal{H}} w_h \log\frac{w_h}{\pi_h}$$

is the Kulback-Leibler divergence between probability distributions w_h and the a priori weights π_h. Since there is no explicit formula for $\tilde{w}_h(Y')$, to simplify their computing, let us replace in (1.11) the distance between the data and the estimate

$$\left\|Y' - \sum_{h\in\mathcal{H}} w_h\hat{\mu}^h(Y)\right\|^2$$

by the following upper bound obtained by the Jensen inequality

$$\left\|Y' - \sum_{h\in\mathcal{H}} w_h\hat{\mu}^h(Y)\right\|^2 \le \sum_{h\in\mathcal{H}} w_h\left\|Y' - \hat{\mu}^h(Y)\right\|^2.$$

Thus we arrive at

$$w_h(Y') = \arg\min_{w_h}\left\{\frac{1}{2\sigma^2}\sum_{h\in\mathcal{H}} w_h\left\|Y' - \hat{\mu}^h(Y)\right\|^2 + \beta K(w_h,\pi_h)\right\}.$$

Now, these weights can be easily computed with a simple algebra

$$w_h(Y') = \frac{\pi_h}{Z(Y')}\exp\left[-\frac{\|\hat{\mu}^h(Y)\|^2 - 2\langle Y',\hat{\mu}^h(Y)\rangle}{2\sigma^2\beta}\right], \qquad (1.12)$$

where

$$Z(Y') = \sum_{h\in\mathcal{H}} \pi_h\exp\left[-\frac{\|\hat{\mu}^h(Y)\|^2 - 2\langle Y',\hat{\mu}^h(Y)\rangle}{2\sigma^2\beta}\right].$$

Obviously, the described above two samples estimation procedure results in loss of a significant statistical information contained in the data. Therefore computing estimates and their aggregation should be done, of course, with the help of the single sample. The main difficulty in implementing this idea is related to the inner product $\langle Y',\hat{\mu}^h(Y)\rangle$ in (1.12). It is clear that we need to replace this inner product by something that should to be close to

it, but depending only on Y. In order to understand how to implement this idea, let us look at the probabilistic structure of $\langle Y', \hat{\mu}^h(Y)\rangle$. We have

$$\begin{aligned}\langle Y', \hat{\mu}^h(Y)\rangle = \sum_{k=1}^n h_k(\mu_k + \sigma\xi'_k)(\mu_k + \sigma\xi_k) = \sum_{k=1}^n h_k\mu_k^2 \\ +\sigma\sum_{k=1}^n h_k\mu_k\xi_k + \sigma\sum_{k=1}^n h_k\mu_k\xi'_k + \sigma^2\sum_{k=1}^n h_k\xi_k\xi'_k.\end{aligned} \tag{1.13}$$

Note that the second line in this equation contains random variables with zero mean for given h. If we want to make use of the single sample, we need first to look at

$$\begin{aligned}\langle Y, \hat{\mu}^h(Y)\rangle = \sum_{k=1}^n h_k(\mu_k + \sigma\xi_k)^2 = \sum_{k=1}^n h_k\mu_k^2 + 2\sigma^2\sum_{k=1}^n h_k \\ +2\sigma\sum_{k=1}^n h_k\mu_k\xi_k + \sigma^2\sum_{k=1}^n h_k(\xi_k^2 - 1).\end{aligned} \tag{1.14}$$

So, we see that the last line in this equation contains also zero mean random variables. Therefore, comparing (1.13) and (1.14), we can conclude that

$$\langle Y', \hat{\mu}^h(Y)\rangle \approx \langle Y, \hat{\mu}^h(Y)\rangle - 2\sigma^2\sum_{k=1}^n h_k$$

in the sense that for any given h

$$\mathbf{E}\langle Y', \hat{\mu}^h(Y)\rangle = \mathbf{E}\langle Y, \hat{\mu}^h(Y)\rangle - 2\sigma^2\sum_{k=1}^n h_k.$$

These intuitive considerations combined with (1.12) result in the exponential weights defined by (1.6)-(1.8).

Notice that in recent years the exponential weighting method has been extensively studied and enough good upper bounds for its performance have been obtained for several statistical models [13, 6, 1, 2].

In this paper, we study this method in the case when $\mathcal{H}$ is a set of ordered multipliers defined as follows :

Definition 1.1. *$\mathcal{H}$ is a set of ordered multipliers if the following properties hold:*

1. *$h_i \in [0,1]$, $i = 1, \ldots, n$ for all $h \in \mathcal{H}$,*

2. $h_{i+1} \le h_i$, $i = 1, \dots, n$ *for all* $h \in \mathcal{H}$,

3. *if for some integer* k *and some* $h, g \in \mathcal{H}$, $h_k < g_k$, *then* $h_i \le g_i$ *for all* $i = 1, \dots, n$.

Property 3 means that vectors in $\mathcal{H}$ may be naturally ordered, since for any $h, g \in \mathcal{H}$ there are only two possibilities $h_i \le g_i$ or $h_i \ge g_i$ for all $i = 1, \dots, n$. Therefore the estimators defined by (1.2), where $\mathcal{H}$ is a set of ordered multipliers, are often called ordered smoothers [11]. Let us emphasize that ordered smoothers are very common in statistics (see for instance [11], [5]).

We start our study of the exponential weighting method with the case $\beta \to 0$. It is easy to see that in this case the limiting estimator has the following form :

$$\bar{\mu}^{\circ}(Y) = h^{\circ}(Y) \cdot Y, \quad \text{where} \quad h^{\circ}(Y) = \arg\min_{h \in \mathcal{H}} \bar{R}(Y, h). \tag{1.15}$$

This standard method of estimate selection with the help of the unbiased risk estimation has long been well known in statistics and goes back to [3, 14]. A classical fact about its performance is given by the following theorem [11].

Theorem 1.1. *Let* $\mathcal{H}$ *be a set of ordered multipliers. Then for any* $\mu \in \mathbb{R}^n$

$$\mathbf{E}_{\mu} \|\bar{\mu}^{\circ}(Y) - \mu\|_2^2 \le r(\mu, \mathcal{H}) + K\sigma^2 \sqrt{1 + \frac{r(\mu, \mathcal{H})}{\sigma^2}}, \tag{1.16}$$

where here and in what follows K *denotes generic constants.*

Equation (1.16) is often called oracle inequality since it controls the risk of $\bar{\mu}^{\circ}(Y)$ via the oracle risk $r(\mu, \mathcal{H})$.

When $\beta > 0$, statistical analysis of the exponential weighting is more involved (see, e.g., [5]). In order to get an oracle inequality similar to (1.16), some additional assumptions about π_h and $\mathcal{H}$ are required.

First, we will make use of a priory weights defined by the following condition

Condition 1.1.

$$\pi_h \stackrel{\text{def}}{=} 1 - \exp\left\{ -\frac{\|h^+\|_1 - \|h\|_1}{\beta} \right\}, \tag{1.17}$$

where $h^+ = \min\{g \in \mathcal{H} : g > h\}$ *and* $\pi_{h^{\max}} = 1$, $h^{\max}$ *is the maximal multiplier in* $\mathcal{H}$.

In this definition and below $\|\cdot\|_1$ stands for ℓ_1-norm in $\mathbb{R}^n$, i.e.,

$$\|h\|_1 = \sum_{i=1}^{n} |h_i|.$$

Along with the above condition, we will need also the following one.

Condition 1.2. *There exists a constant $K_\circ > 0$ such that*

$$\|h\|^2 - \|g\|^2 \geq K_\circ\big(\|h\|_1 - \|g\|_1\big) \tag{1.18}$$

for all $h \geq g$ from $\mathcal{H}$.

An upper bound for mean square risk of the exponential weighting method is given by the following theorem [5].

Theorem 1.2. *Assume that $\mathcal{H}$ is a set of ordered multipliers, $\beta \geq 4$, and Conditions 1.1, 1.2 hold. Then, uniformly in $\mu \in \mathbb{R}^n$,*

$$\mathbf{E}_\mu\|\bar{\mu}^\beta(Y) - \mu\|^2 \leq r^\beta(\mu, \mathcal{H}) + \sigma^2 C(K_\circ, \beta), \tag{1.19}$$

where

$$r^\beta(\mu, \mathcal{H}) \stackrel{\text{def}}{=} r(\mu, \mathcal{H}) + 2\beta\sigma^2 \log\left[2 + \frac{r(\mu, \mathcal{H})}{\sigma^2}\right]. \tag{1.20}$$

Here and in what follows $C(K_\circ, \beta)$ denotes strictly positive and bounded constants depending on $K_\circ$ and β.

At first glance, it may seem comparing the remainder terms in Equations (1.16) and (1.19)-(1.20), that the exponential weighting should be significantly better than the classical methods of model selection. In fact, this is not exactly so, and to find out what is really going on, we need to understand first of all what methods were used in proving Theorems 1.1 and 1.2.

Theorem 1.1 results from the following concentration inequality (see Kneip [11]) :

$$\mathbf{P}_\mu\Big\{\|\bar{\mu}^\circ(Y) - \mu\| \geq \sqrt{r(\mu, \mathcal{H})} + K\sigma + x\Big\} \leq \exp\Big\{-\frac{x^2}{K\sigma^2}\Big\}, \ x \geq 0. \tag{1.21}$$

At the same time, Theorem 1.2 represents a fact of an entirely different class. At the very core in the proof of (1.19) in [5] is an original method proposed in [13] which is based on Stein's formula [18] for the unbiased risk estimate. In essence, this means that in contrast to (1.21) Equation (1.19)

cannot control the deviations of the square losses $\|\bar{\mu}^\beta(Y) - \mu\|^2$ from the oracle risk.

Note also that Theorem 1.2 holds true only for $\beta \geq 4$. Unfortunately, it is impossible to say whether this condition is crucial for practical applications of the exponential weighting, or it is a purely mathematical constraint resulting from the method of the proof.

In order to understand more profoundly statistical properties of the exponential weighting method, we focus in this paper on the so-called excess risk

$$\Delta^\beta(Y, \mathcal{H}) \stackrel{\text{def}}{=} \big[\|\bar{\mu}^\beta(Y) - \mu\|^2 - r^\beta(\mu, \mathcal{H})\big]_+,$$

where $[x]_+ = \max(0, x)$ and $r^\beta(\mu, \mathcal{H})$ is defined by (1.20).

The next theorem, controlling moments of $\Delta^\beta(Y, \mathcal{H})$, is the main result of this paper.

Theorem 1.3. *Assume that $\mathcal{H}$ is a set of ordered multipliers and Conditions 1.1, 1.2 hold. Then, uniformly in $\mu \in \mathbb{R}^n$ and in $m \geq 1$,*

$$\mathbf{E}_\mu^{1/m}\big[\Delta^\beta(Y, \mathcal{H})\big]_+^m \leq K\sigma\sqrt{mr^\beta(\mu, \mathcal{H})} + K\sigma^2 m + C(K_\circ, \beta)\sigma^2. \tag{1.22}$$

Notice that compared to majority of results related to the exponential weighting, see e.g. [13, 17, 6, 5], the main advantage of this theorem is that it holds for any $\beta > 0$ and provides an upper bound for the excess risk that does depend neither n no the cardinality of $\mathcal{H}$. It justifies, in particular, the use of the exponential weighting with $\beta = 1$ that demonstrates a good performance in practice as shown in the next section.

Combining Theorem 1.3 with the Markov inequality, one obtains the following fact similar to (1.21).

Theorem 1.4. *Assume that $\mathcal{H}$ is a set of ordered multipliers and Conditions 1.1, 1.2 hold. Then, uniformly in $\mu \in \mathbb{R}^n$ and $x \geq \sigma$,*

$$\mathbf{P}_\mu\Big\{\|\bar{\mu}^\beta(Y) - \mu\| \geq \sqrt{r^\beta(\mu, \mathcal{H})} + x\Big\} \leq \exp\Big\{-\frac{x^2}{K\sigma^2} + C(K_\circ, \beta)\Big\}.$$

2 Simulation study

To illustrate numerically Theorem 1.3, the following experiment has been carried out. Its goal was to find out how the exponential weighting with $\beta = \{0, 1, 2, 4\}$ works in regression estimation with the help of the cubic

smoothing splines. Recall that these splines are usually used in recovering an unknown smooth function $f(x)$, $x \in [0,1]$ from the noisy observations

$$Y_i' = f(X_i) + \epsilon\xi_i', \quad i = 1, \dots, n, \tag{2.1}$$

where it is assumed that the design points X_i belong to $(0,1)$ and ξ' is a standard white Gaussian noise. Smoothing spline of order $2m-1$ is defined by

$$\hat{f}_\alpha(x, X, Y') = \arg\min_f \left\{ \frac{1}{n} \sum_{i=1}^{n} [Y_i' - f(X_i)]^2 + \alpha \int_0^1 [f^{(m)}(u)]^2 \, du \right\}, \tag{2.2}$$

where $f^{(m)}(\cdot)$ denotes the derivative of order m and $\alpha \in \mathcal{A} \subset \mathbb{R}^+$ is a smoothing parameter. In what follows the set of possible smoothing parameters $\mathcal{A}$ is assumed to be finite.

In order to show that the regression estimation with the help of the smoothing splines is equivalent to the sequence space model described by (1.1) and (1.2), consider the Demmler-Reinsch [7] basis $\psi_k(x)$, $x \in [0,1]$, $k = 1, \dots, n$ having double orthogonality

$$\langle \psi_k, \psi_l \rangle_n = \delta_{kl}, \quad \int_0^1 \psi_k^{(m)}(x) \psi_l^{(m)}(x) \, dx = \delta_{kl} \lambda_k, \ k, l = 1, \dots, n, \tag{2.3}$$

where $\delta_{kl} = 1$ if $k = l$, and $\delta_{kl} = 0$ otherwise. Here and below $\langle u, v \rangle_n$ stands for the inner product

$$\langle u, v \rangle_n = \frac{1}{n} \sum_{i=1}^{n} u(X_i) v(X_i).$$

Let us assume for definiteness that the eigenvalues λ_k are sorted in ascending order $\lambda_1 \le \dots \le \lambda_n$.

With this basis, representing the underlying regression function as follows:

$$f(x) = \sum_{k=1}^{n} \psi_k(x) \mu_k, \tag{2.4}$$

we get from (2.1) and (2.3)

$$Y_k = \frac{1}{n} \sum_{i=1}^{n} Y_i' \psi_k(X_i) = \mu_k + \frac{\varepsilon}{\sqrt{n}} \xi_k, \tag{2.5}$$

where $\mu_k = \langle f, \psi_k \rangle_n$ and ξ is a standard white Gaussian noise. So, substituting (2.4) in (2.2) and using (2.3), we arrive at

$$\hat{f}_\alpha(x, X, Y') = \sum_{k=1}^{n} \hat{\mu}_k \psi_k(x),$$

where

$$\hat{\mu} = \arg\min_{\mu} \left\{ \sum_{k=1}^{n} [Y_k - \mu_k]^2 + \alpha \sum_{k=1}^{n} \lambda_k \mu_k^2 \right\}.$$

It can be seen easily that

$$\hat{\mu}_k = \frac{Y_k}{1 + \alpha \lambda_k}$$

and thus the spline regression model (2.1)-(2.2) is equivalent to the sequence space model defined by (1.1) and (1.2) with $\sigma = \varepsilon/\sqrt{n}$ and

$$\mathcal{H} = \left\{ h : \ h_k = \frac{1}{1 + \alpha \lambda_k}, \ k = 1, \dots, n, \ \alpha \in \mathcal{A} \right\}. \tag{2.6}$$

In order to simulate cubic splines, the following family of ordered multipliers was used

$$\mathcal{H} = \left\{ h : h_k = \frac{1}{1 + [\alpha \pi (k-1)]^4}, \ k = 1, \dots, n, \ \alpha \in \mathcal{A}^n \right\},$$

where

$$\mathcal{A}^n = \left\{ \alpha : \alpha = (1 + \varepsilon)^s, \ s = 0, 1, \dots, \lfloor \log(n) / \log(1 + \varepsilon) \rfloor \right\}$$

with $\varepsilon = 0.01$. The motivation of this family is due to the well-known asymptotic formula $\lambda_k \asymp (\pi k)^4$, $k \gg 1$ (see [7]).

The simulations were organized as follows. For given $A \in [0, 50]$, 40000 replications of the observations

$$Y_k = \mu_k(A) + \xi_k, \ k = 1, \dots, 300$$

were generated. Here $\mu(A) \in \mathbb{R}^{300}$ is a Gaussian vector with independent components and

$$\mathbf{E}\mu_k(A) = 0, \quad \mathbf{E}\mu_k^2(A) = A \exp\left(-\frac{k^2}{2\Sigma^2} \right).$$

Next, the mean oracle risk

$$\bar{r}(A, \mathcal{H}) = \mathbf{E} \min_{h \in \mathcal{H}} \left\{ \|(1 - h) \cdot \mu(A)\|^2 + \|h\|^2 \right\}$$

and the absolute excess risk

$$\bar{\Delta}_1^{\beta}(A) = \mathbf{E}\Big[\|\mu(A) - \bar{\mu}^{\beta}(Y)\|^2 - \bar{r}(A,\mathcal{H})\Big]_+,$$

were computed with the help of the Monte-Carlo method. Finally, the data $\big\{\bar{r}(A,\mathcal{H}),\ \bar{\Delta}_1^{\beta}(A),\ A \in [0,50],\ \beta = 0,1,2,4\big\}$ are plotted on Figure 1 to illustrate graphically statistical properties of the exponential weighting method.

Looking at this picture we see that there is no universal β minimizing the excess risk uniformly in μ, but it seems that a reasonable choice would be $\beta \approx 1$. Note also that the exponential weighting with $\beta \in [1,4]$ works better compared to the classical unbiased risk estimation ($\beta = 0$) when $r(\mu,\mathcal{H})/\sigma^2$ is not large.

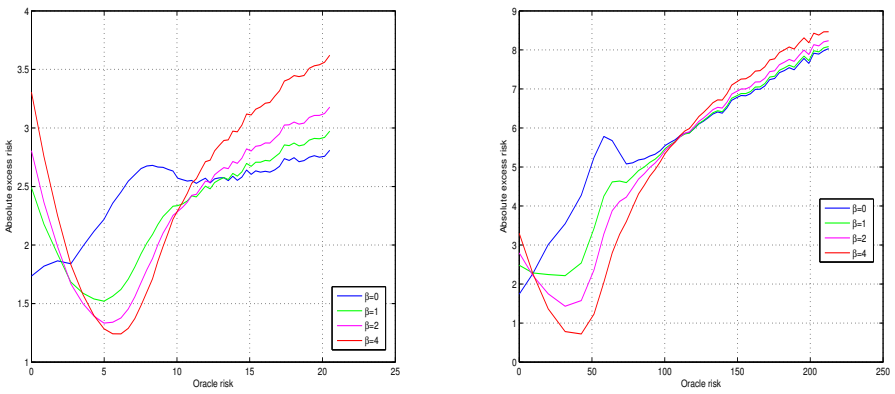

Figure 1: Absolute excess risks $\bar{\Delta}_1^{\beta}(\cdot)$ (left panel $\Sigma = 5$; right panel $\Sigma = 50$).

3 Proofs

3.1 Auxiliary facts

In what follows it is assumed that $\mathcal{H}$ be a set of ordered multipliers. Basic probabilistic facts related to the ordered multipliers are described is the following lemma.

Lemma 3.1. *Let ξ_i be i.i.d. $\mathcal{N}(0,1)$. Then for any $\alpha \in (0, 1/4)$*

$$\mathbf{P}\Big\{\max_{h\in\mathcal{H}}\Big[\pm\sum_{i=1}^{n} h_i(\xi_i^2-1) - \alpha\sum_{i=1}^{n} h_i^2\Big] \geq \frac{x}{\alpha}\Big\} \leq \exp\Big(-\frac{x}{K}\Big), \tag{3.1}$$

$$\mathbf{P}\Big\{\max_{h\in\mathcal{H}}\Big[\sum_{i=1}^{n}(1-h_i)\xi_i\mu_i - \alpha\sum_{i=1}^{n}(1-h_i)^2\mu_i^2\Big] \geq \frac{x}{\alpha}\Big\} \leq \exp\Big(-\frac{x}{K}\Big), \tag{3.2}$$

where K is a generic constant.

The proof of this lemma follows immediately from Lemma 2 in [8].

Lemma 3.2. *I) If for some $R, r \geq 0$*

$$R \leq r + \min_{\alpha\in(0,\epsilon]}\Big\{\frac{x^2}{\alpha} + \alpha R\Big\}, \quad \textit{then} \quad \sqrt{R} \leq \sqrt{r} + |x|\max\Big\{2, \sqrt{\frac{2}{\epsilon}}\Big\}.$$

II) If for some $R, r \geq 0$

$$R \leq r + \min_{\alpha\in(0,\epsilon)}\Big\{\frac{x^2}{\alpha} + \alpha r\Big\}, \quad \textit{then} \quad \sqrt{R} \leq \sqrt{r} + |x|\max\Big\{1, \sqrt{\frac{2}{\epsilon}}\Big\}.$$

Proof. It is easy to check with a simple algebra that

$$\min_{\alpha\in(0,\epsilon]}\Big\{\frac{x^2}{\alpha} + \alpha R\Big\} = 2|x|\sqrt{R}\mathbf{1}\big\{x \leq \epsilon\sqrt{R}\big\} + (\epsilon^{-1}x^2 + \epsilon R)\mathbf{1}\big\{x > \epsilon\sqrt{R}\big\}$$
$$2|x|\sqrt{R}\mathbf{1}\big\{x \leq \epsilon\sqrt{R}\big\} + 2\epsilon^{-1}x^2\mathbf{1}\big\{x > \epsilon\sqrt{R}\big\}.$$

Therefore if $x > \epsilon\sqrt{R}$, then

$$R \leq r + 2\epsilon^{-1}x^2,$$

and so,

$$\sqrt{R} \leq \sqrt{r} + \sqrt{2}|x|/\sqrt{\epsilon}.$$

Next, when $x \leq \epsilon\sqrt{R}$ we have

$$R \leq r + 2|x|\sqrt{R},$$

or, equivalently,

$$(\sqrt{R} - |x|)^2 \leq r + x^2.$$

Hence

$$\sqrt{R} \leq \sqrt{r} + 2|x|.$$

The proof of the second part of the lemma is quite similar. ■

Lemma 3.3. *Let ζ be a nonnegative random variable with $\mathbf{E}\zeta^m < \infty$. Then for any $A > 0$*

$$\mathbf{E}^{1/m} \log^m(A + \zeta) \le \log(A + \mathbf{E}^{1/m}\zeta^m).$$

Proof. Consider the following function

$$F(z) = \max_{x > -A} \big\{\log(A + x) - zx\big\},\ z \in \mathbb{R}^+.$$

It can be checked with a simple algebra that

$$F(z) = -\log(z) + Az - 1.$$

and therefore for any $z \ge 0$

$$\log(A + x) \le zx + F(z).$$

So, we have

$$\mathbf{E}^{1/m} \log^m(A + \zeta) \le z\mathbf{E}^{1/m}\zeta^m + F(z)$$

and minimizing the right-hand side at this equation in $z \ge 0$, we complete the proof. ■

Lemma 3.4. *Let*

$$\begin{aligned} \hat{h}^\epsilon = \max\Big\{ & h \in \mathcal{H} : \bar{R}(Y, h) - \bar{R}(Y, h^\circ) \\ & \le 2\beta\epsilon\sigma^2 \Big[\|h\|^2 - \|h^\circ\|^2 + \epsilon^{-1}\Big]\Big\}, \end{aligned} \tag{3.3}$$

where $\epsilon \in (0, 1/(2\beta))$ and $h^\circ(Y)$ is defined by (1.15). Then for any integer $m \ge 1$

$$\sqrt{1 - 2\beta\epsilon}\mathbf{E}_\mu^{1/(2m)}\|\hat{h}^\epsilon\|^{2m} \le \sqrt{\frac{r(\mu, \mathcal{H})}{\sigma^2}} + K\sqrt{(m + \beta)}. \tag{3.4}$$

Proof. By the definition of $\bar{R}(Y, h)$ (see (1.8))) and (3.3), we get

$$\begin{aligned} \hat{h}^\epsilon = \max\Big\{ & h : \|(1 - h) \cdot \mu\|^2 + \sigma^2(1 - 2\beta\epsilon)\|h\|^2 \\ & + 2\sigma \sum_{i=1}^n (1 - h_i)^2 \mu_i \xi_i + \sigma^2 \sum_{i=1}^n (h_i^2 - 2h_i)(\xi_i^2 - 1) \\ & \le \|(1 - \hat{h}) \cdot \mu\|^2 + \sigma^2(1 - 2\beta\epsilon)\|\hat{h}\|^2 \\ & + 2\sigma \sum_{i=1}^n (1 - \hat{h}_i)^2 \mu_i \xi_i + \sigma^2 \sum_{i=1}^n (\hat{h}_i^2 - 2\hat{h}_i)(\xi_i^2 - 1) + 2\beta\sigma^2 \Big\}. \end{aligned}$$

Let us fix some $\gamma, \gamma' \in (0, 1/4)$. Then we can rewrite the above equation as follows:

$$\begin{aligned}\hat{h}^\epsilon = \max\Big\{ h \in \mathcal{H} :\ & \sigma^2(1-2\beta\epsilon-\gamma)\|h\|^2 + 2\sigma\sum_{i=1}^n (1-h_i)^2\mu_i\xi_i \\ & + \|(1-h)\cdot\mu\|^2 + \sigma^2\sum_{i=1}^n (h_i^2-2h_i)(\xi_i^2-1) + \gamma\sigma^2\|h\|^2 \\ & \le (1+\gamma')\|(1-\hat{h})\cdot\mu\|^2 + \sigma^2(1-2\beta\epsilon+\gamma')\|\hat{h}\|^2 \\ & + 2\sigma\sum_{i=1}^n (1-\hat{h}_i)^2\mu_i\xi_i - \gamma'\|(1-\hat{h})\cdot\mu\|^2 \\ & + \sigma^2\sum_{i=1}^n (\hat{h}_i^2-2\hat{h}_i)(\xi_i^2-1) - \gamma'\sigma^2\|\hat{h}\|^2 + 2\beta\sigma^2 \Big\}.\end{aligned}$$

Hence

$$\begin{aligned}\hat{h}^\epsilon \le \tilde{h}^\epsilon \stackrel{\text{def}}{=} \max\Big\{ h \in \mathcal{H} :\ & \sigma^2(1-2\beta\epsilon-\gamma)\|h\|^2 \\ & + \min_{g\in\mathcal{H}}\Big[2\sigma\sum_{i=1}^n (1-g_i)^2\mu_i\xi_i + \|(1-g)\cdot\mu\|^2\Big] \\ & + \min_{g\in\mathcal{H}}\Big[\sigma^2\sum_{i=1}^n (g_i^2-2g_i)(\xi_i^2-1) + \gamma\sigma^2\|g\|^2\Big] \\ & \le (1+\gamma')\|(1-h^\circ)\cdot\mu\|^2 + (1-2\beta\epsilon+\gamma')\sigma^2\|h^\circ\|^2 \\ & + \max_{g\in\mathcal{H}}\Big[2\sigma\sum_{i=1}^n (1-g_i)^2\mu_i\xi_i - \gamma'\|(1-g)\cdot\mu\|^2\Big] \\ & + \max_{g\in\mathcal{H}}\Big[\sigma^2\sum_{i=1}^n (g_i^2-2g_i)(\xi_i^2-1) - \gamma'\sigma^2\|g\|^2\Big] + 2\beta\sigma^2 \Big\}.\end{aligned} \tag{3.5}$$

Next, bounding *max* and *min* in (3.5) with the help of Lemma 3.1, we obtain that for any integer $m > 0$ and any $\gamma, \gamma' \in (0, 1/4)$ the following inequality holds

$$\begin{aligned}(1-2\beta\epsilon-\gamma)\sigma^2\mathbf{E}_\mu^{1/m}\|\tilde{h}^\epsilon\|^{2m} \le (1+\gamma')\mathbf{E}_\mu^{1/m}R^m(\mu,h^\circ) \\ + \frac{K(m!)^{1/m}\sigma^2}{\gamma'} + \frac{K[(m!)^{1/m}+\beta]\sigma^2}{\gamma},\end{aligned}$$

where $R(\cdot,\cdot)$ is defined by (1.3).

Therefore, applying Lemma 3.2, we obtain from the above equation

$$\sqrt{1-2\beta\epsilon}\sigma\mathbf{E}_\mu^{1/(2m)}\|\tilde{h}^\epsilon\|^{2m} \le \mathbf{E}_\mu^{1/(2m)}R^m(\mu,h^\circ) + K\sqrt{m+\beta}\sigma. \tag{3.6}$$

In order to control the expectation at the right-hand side in (3.6), note that the following inequality

$$\sum_{i=1}^n [1-h_i^\circ]^2 Y_i^2 + 2\sigma^2\sum_{i=1}^n h_i^\circ \le \sum_{i=1}^n [1-g_i]^2 Y_i^2 + 2\sigma^2\sum_{i=1}^n g_i$$

holds for any given $g \in \mathcal{H}$. Hence, we have

$$\begin{aligned} R(\mu,h^\circ) &+ 2\sigma\sum_{i=1}^n (1-h_i^\circ)^2\mu_i\xi_i + \sigma^2\sum_{i=1}^n (h_i^\circ)^2 - 2h_i^\circ)(\xi_i^2-1) \\ &\le R(\mu,g) + 2\sigma\sum_{i=1}^n (1-g_i)^2\mu_i\xi_i + \sigma^2\sum_{i=1}^n (g_i^2-2g_i)(\xi_i^2-1). \end{aligned}$$

So, for any $\gamma,\gamma' \in (0,1/4]$, we get with this equation and Lemma 3.1

$$\begin{aligned} \mathbf{E}_\mu^{1/m}R^m(\mu,h^\circ) &\le R(\mu,g) + \gamma\mathbf{E}_\mu^{1/m}R^m(\mu,h^\circ) + \gamma' R(\mu,g) \\ &+ 2\sigma\mathbf{E}^{1/m}\max_{h\in\mathcal{H}}\left[-\sum_{i=1}^n (1-h_i)^2\mu_i\xi_i - \frac{\gamma}{2\sigma}\sum_{i=1}^n (1-h_i)^2\mu_i^2\right]^m \\ &+ \sigma^2\mathbf{E}^{1/m}\max_{h\in\mathcal{H}}\left[\sum_{i=1}^n (2h_i-h_i^2)(\xi_i^2-1) - \gamma\sum_{i=1}^n h_i^2\right]^m \\ &+ 2\sigma\mathbf{E}^{1/m}\max_{h\in\mathcal{H}}\left[-\sum_{i=1}^n (1-h_i)^2\mu_i\xi_i - \frac{\gamma'}{2\sigma}\sum_{i=1}^n (1-h_i)^2\mu_i^2\right]^m \\ &+ \sigma^2\mathbf{E}^{1/m}\max_{h\in\mathcal{H}}\left[\sum_{i=1}^n (2h_i-h_i^2)(\xi_i^2-1) - \gamma'\sum_{i=1}^n h_i^2\right]^m \\ &\le R(\mu,g) + \frac{K\sigma^2 m}{\gamma} + \gamma\mathbf{E}_\mu^{1/m}R^m(\mu,h^\circ) + \frac{K\sigma^2 m}{\gamma'} + \gamma' R(\mu,g). \end{aligned}$$

Next, minimizing the right-hand side in $g \in \mathcal{H}$ and applying Lemma 3.2, we obtain

$$\mathbf{E}_\mu^{1/(2m)}R^m(\mu,h^\circ) \le \sqrt{r(\mu,\mathcal{H})} + K\sigma\sqrt{m}$$

and substituting this inequality in (3.6), we complete the proof. ■

The next lemma collects some useful facts about a priory weights defined by (1.17). Let

$$\mathcal{D}_h = \big\{g \in \mathcal{H} : \|h\|_1 \le \|g\|_1 \le \|h\|_1 + 1\big\}. \tag{3.7}$$

Lemma 3.5. *Under Condition 1.1, for any $h \in \mathcal{H}$, the following assertions hold:*

$$\sum_{g \ge h} \pi_g \exp\Big\{-\frac{\|g\|_1}{\beta}\Big\} = \exp\Big\{-\frac{\|h\|_1}{\beta}\Big\}, \tag{3.8}$$

$$\sum_{g \le h} \pi_g \le 1 + \frac{\|h\|^2}{K_\circ \beta}, \tag{3.9}$$

$$\sum_{g \in \mathcal{D}_h} \pi_g \le 1 + \frac{1}{\beta}, \tag{3.10}$$

$$\sum_{g \in \mathcal{D}_h} \pi_g \ge \frac{1}{2\beta} \exp\Big(-\frac{1}{\beta}\Big). \tag{3.11}$$

The proof of this lemma can be found in [5].

The following technical lemma, whose proof is given in [5], is a cornerstone in the proof of Theorem 1.3.

Lemma 3.6. *Suppose $\{q_h \le 1,\ h \in \mathcal{H}\}$ is a nonnegative sequence such that for all $h \ge \tilde{h}$*

$$q_h \le \exp\Big\{-\gamma\big[\|h\|_1 - \|\tilde{h}\|_1\big] - 1\Big\}, \quad \gamma > 0.$$

Let

$$W_h = \pi_h q_h \Big(\sum_{g \in \mathcal{H}} \pi_g q_g\Big)^{-1}$$

and $\mathcal{G}$ be a subset in $\mathcal{H}$. Then

$$\sum_{h \in \mathcal{H}} W_h \log \frac{\pi_h}{W_h} \le \log\Big\{\sum_{h \le \tilde{h}} \pi_h + \exp\big[R(\mathcal{G}, \gamma)\big]\Big\},$$

where

$$E(\mathcal{G}, \gamma) = \log\Big(\frac{2}{\gamma\beta \mathrm{e}} + \sum_{h \in \mathcal{G}} \pi_h\Big) + \Big(\sum_{h \in \mathcal{G}} \pi_h q_h\Big)^{-1} \Big(\frac{8}{\gamma\beta \mathrm{e}} + \sum_{h \in \mathcal{G}} \pi_h\Big). \tag{3.12}$$

3.2 Proof of Theorem 1.3

The proof is based essentially on the approach proposed in [9, 5]. We begin with a basic inequality for the excess risk. By Jensen's inequality we have

$$\begin{aligned}\|\bar{\mu}^{\beta}(Y)-\mu\|^2 &= \Big\|\sum_{h\in\mathcal{H}} w_h(Y)\big(h\cdot Y-\mu\big)\Big\|^2 \\ &\le \sum_{h\in\mathcal{H}} w_h(Y)\big\|h\cdot Y-\mu\big\|^2.\end{aligned} \tag{3.13}$$

This is why in what follows we will deal with $\sum_{h\in\mathcal{H}} w_h(Y)\big\|h\cdot Y-\mu\big\|^2$. In order to bound this value from above, let us express $\big\|h\cdot Y-\mu\big\|^2$ in terms of the unbiased risk estimate of $h\cdot Y$. With a simple algebra we get

$$\begin{aligned}\big\|h\cdot Y-\mu\big\|^2 &= \big\|h\cdot Y-Y+\sigma\xi\big\|^2 \\ &= \big\|(1-h)\cdot Y\big\|^2+\sigma^2\big\|\xi\big\|^2-2\sigma\langle\xi,(1-h)\cdot(\mu+\sigma\xi)\rangle \\ &= \big\|(1-h)\cdot Y\big\|^2+2\sigma^2\sum_{i=1}^n h_i-\sigma^2\big\|\xi\big\|^2 \\ &\quad -2\sigma\langle\xi,(1-h)\cdot\mu\rangle+2\sigma^2\sum_{i=1}^n h_i(\xi_i^2-1).\end{aligned} \tag{3.14}$$

Denote for brevity

$$\begin{aligned}\tilde{R}(Y,h) &\stackrel{\text{def}}{=} \big\|(1-h)\cdot Y\big\|^2+2\sigma^2\sum_{i=1}^n h_i-\sigma^2\big\|\xi\big\|^2, \\ \tilde{r}(Y,\mathcal{H}) &\stackrel{\text{def}}{=} \min_{h\in\mathcal{H}}\bar{R}(Y,h), \quad h^*=\arg\min_{h\in\mathcal{H}}\mathbf{E}\bar{R}(Y,h).\end{aligned} \tag{3.15}$$

With these notations and (3.14), we obtain

$$\begin{aligned}\sum_{h\in\mathcal{H}} w_h(Y)\big\|h\cdot Y-\mu\big\|^2 &= \tilde{r}(Y,\mathcal{H})+\sum_{h\in\mathcal{H}} w_h(Y)\big[\tilde{R}(Y,h)-\tilde{r}(Y,\mathcal{H})\big] \\ &\quad+\sum_{h\in\mathcal{H}} w_h(Y)\Big[2\sigma^2\sum_{i=1}^n h_i(\xi_i^2-1)-2\sigma\sum_{i=1}^n(1-h_i)\xi_i\mu_i\Big].\end{aligned} \tag{3.16}$$

It is also obvious that

$$\begin{aligned}\tilde{r}(Y,\mathcal{H}) \le \tilde{R}(Y,h^*) &= \|(1-h^*)\cdot\mu\|^2 + \sigma^2\|h^*\|^2 \\ &+ \sigma^2\sum_{i=1}^{n}(h_i^{*2}-2h_i^*)(\xi_i^2-1) + 2\sigma\sum_{i=1}^{n}(1-h_i^*)^2\mu_i\xi_i \\ &= r(\mu,\mathcal{H}) + \sigma^2\sum_{i=1}^{n}(h_i^{*2}-2h_i^*)(\xi_i^2-1) + 2\sigma\sum_{i=1}^{n}(1-h_i^*)^2\mu_i\xi_i.\end{aligned} \tag{3.17}$$

Recalling the definition of the weights $w_h(Y)$, we rewrite them as follows:

$$w_h(Y) = \pi_h \exp\left[-\frac{\tilde{R}(Y,h)-\tilde{r}(Y,\mathcal{H})}{2\beta\sigma^2}\right] \Big/ \sum_{g\in\mathcal{H}}\pi_g \exp\left[-\frac{\tilde{R}(Y,g)-\tilde{r}(Y,\mathcal{H})}{2\beta\sigma^2}\right]$$

and we obtain

$$\begin{aligned}\tilde{R}(Y,h)-\tilde{r}(Y,\mathcal{H}) &= 2\beta\sigma^2\sum_{h\in\mathcal{H}} w_h(Y)\log\frac{\pi_h}{w_h(Y)} \\ &- 2\beta\sigma^2\log\left\{\sum_{h\in\mathcal{H}}\pi_h\exp\left[-\frac{\tilde{R}(Y,h)-\tilde{r}(Y,\mathcal{H})}{2\beta\sigma^2}\right]\right\}.\end{aligned} \tag{3.18}$$

Therefore, substituting Equations (3.17) and (3.18) in (3.16), we arrive at the basic inequality

$$\begin{aligned}\sum_{h\in\mathcal{H}} w_h(Y)\left\|h\cdot Y-\mu\right\|^2 &\le r(\mu,\mathcal{H}) \\ &+ \delta_1(\mu)+\delta_2(\mu)+2\beta\sigma^2\delta_3(\mu)+2\beta\sigma^2\delta_4(\mu),\end{aligned} \tag{3.19}$$

where

$$\begin{aligned}\delta_1(\mu) =& \sigma^2\sum_{i=1}^{n}(h_i^{*2}-2h_i^*)(\xi_i^2-1) + 2\sigma\sum_{i=1}^{n}(1-h_i^*)^2\mu_i\xi_i, \\ \delta_2(\mu) =& \sum_{h\in\mathcal{H}} w_h(Y)\left[2\sigma^2\sum_{i=1}^{n}h_i(\xi_i^2-1) - 2\sigma\sum_{i=1}^{n}(1-h_i)\xi_i\mu_i\right], \\ \delta_3(\mu) =& -\log\left\{\sum_{h\in\mathcal{H}}\pi_h\exp\left[-\frac{\tilde{R}(Y,h)-\tilde{r}(Y,\mathcal{H})}{2\beta\sigma^2}\right]\right\}, \\ \delta_4(\mu) =& \sum_{h\in\mathcal{H}} w_h(Y)\log\frac{\pi_h}{w_h(Y)}.\end{aligned}$$

The first term at the right-hand side of (3.19) can be easily controlled. Indeed, denote for brevity one can check easily that

$$\begin{aligned}\mathbf{E}\delta_1^2(\mu) &= 2\sigma^4\sum_{i=1}^n(h_i^{*2}-2h_i^*)^2+4\sigma^2\sum_{i=1}^n(1-h_i^*)^4\mu_i^2\\ &\le \sigma^2\Big[8\sigma^2\sum_{i=1}^n h_i^{*2}+4\sum_{i=1}^n(1-h_i^*)^2\mu_i^2\Big]\le 8\sigma^2 r(\mu,\mathcal{H}).\end{aligned}\tag{3.20}$$

Next, since by the definition of ordered multipliers $(h_i^*)^2-2h_i^*\le 0$, we get by a simple algebra that for any $\lambda>0$

$$\begin{aligned}\log[\exp(\lambda\delta_1(\mu))] &= \sum_{i=1}^n\Big\{\lambda\sigma^2(2h_i^*-h_i^{*2})\\ &-\frac{1}{2}\log\Big[1+2\lambda\sigma^2(2h_i^{*2}-h_i^*)\Big]\frac{2\lambda^2\sigma^2(1-h_i^*)^4\mu_i^2}{1+2\lambda\sigma^2(2h_i^{*2}-h_i^*)}\Big\}\\ &\le \sum_{i=1}^n\Big\{\lambda^2\sigma^4(2h_i^{*2}-h_i^*)^2+2\lambda^2\sigma^2(1-h_i^*)^4\mu_i^2\Big\}\\ &\le 4\lambda^2\sigma^2 r(\mu,\mathcal{H}).\end{aligned}\tag{3.21}$$

Therefore, combining (3.20) and (3.21), we obtain that for any $\lambda>0$

$$\mathbf{E}\exp\Big\{\frac{\lambda\delta_1(\mu)}{\sigma\sqrt{8r(\mu,\mathcal{H})}}\Big\}\le\exp\Big(\frac{\lambda^2}{2}\Big).\tag{3.22}$$

It follows immediately from convexity of the function $\exp(\cdot)$ that

$$z\le\exp(z-1),$$

and so, for any $x,\ \lambda>0$

$$x^m\le\frac{\exp(\lambda mx-m)}{\lambda^m}.$$

Thus, by (3.22)

$$\mathbf{E}\Big[\frac{\delta_1(\mu)}{\sigma\sqrt{8r(\mu,\mathcal{H})}}\Big]_+^m\le\exp\Big\{\frac{\lambda^2m^2}{2}-m\log(\lambda)-m\Big\}.$$

Minimizing the right-hand side at this equation in $\lambda>0$, we arrive at

$$\mathbf{E}\Big[\frac{\delta_1(\mu)}{\sigma\sqrt{8r(\mu,\mathcal{H})}}\Big]_+^m\le\exp\Big\{-\frac{m}{2}+\frac{m}{2}\log(m)\Big\}$$

and thus

$$\mathbf{E}^{1/m}[\delta_1(\mu)]_+^m \le \sigma\sqrt{Kmr(\mu,\mathcal{H})}. \tag{3.23}$$

The last three terms at the right-hand side in (3.19) are bounded from above with the help of more delicate arguments. Since $\sum_{h\in\mathcal{H}} w_h(Y) = 1$, we obtain that for any $\gamma \in (0, 1/4)$

$$\begin{aligned}
\delta_2(\mu) \le & 2\sigma\left[\sum_{h\in\mathcal{H}} w_h(Y)\left(\sigma\sum_{i=1}^n h_i(\xi_i^2-1) - \sigma\gamma\sum_{i=1}^n h_i^2\right)\right]_+ \\
& + 2\sigma\left[\sum_{h\in\mathcal{H}} w_h(Y)\left(\sum_{i=1}^n (1-h_i)\xi_i\mu_i - \frac{\gamma}{\sigma}\sum_{i=1}^n (1-h_i)^2\mu_i^2\right)\right]_+ \\
& + 2\gamma\sum_{h\in\mathcal{H}} w_h(Y)R(\mu,h) \\
\le & 2\sigma^2 \max_{h\in\mathcal{H}}\left[\sum_{i=1}^n h_i(\xi_i^2-1) - \gamma\sum_{i=1}^n h_i^2\right]_+ \\
& + 2\sigma\max_{h\in\mathcal{H}}\left[\sum_{i=1}^n (1-h_i)\xi_i\mu_i - \frac{\gamma}{\sigma}\sum_{i=1}^n (1-h_i)^2\mu_i^2\right]_+ \\
& + 2\gamma\sum_{h\in\mathcal{H}} w_h(Y)R(\mu,h).
\end{aligned}$$

Hence, using Lemma 3.1, we obtain from this equation

$$\mathbf{E}_\mu^{1/m}[\delta_2(\mu)]_+^m \le 2\gamma\mathbf{E}_\mu^{1/m}\left[\sum_{h\in\mathcal{H}} w_h(Y)R(\mu,h)\right]^m + \frac{Km\sigma^2}{\gamma} \tag{3.24}$$

Our next step is to relate

$$\sum_{h\in\mathcal{H}} w_h(Y)R(\mu,h) \quad \text{and} \quad \sum_{h\in\mathcal{H}} w_h(Y)\|\mu - h\cdot Y\|^2.$$

Since

$$\begin{aligned}
\sum_{h\in\mathcal{H}} w_h(Y)\|\mu - h\cdot Y\|^2 = & \sum_{h\in\mathcal{H}} w_h(Y)R(\mu,h) \\
& + 2\sigma\sum_{h\in\mathcal{H}} w_h(Y)\Big[(1-h_i) - (1-h_i)^2\Big]\mu_i\xi_i + \sigma^2\sum_{h\in\mathcal{H}} w_h(Y)h_i^2(\xi_i^2-1),
\end{aligned}$$

we get

$$\sum_{h\in\mathcal{H}} w_h(Y)R(\mu,h) \le \sum_{h\in\mathcal{H}} w_h(Y)\|\mu - h\cdot Y\|^2 + 2\sigma\sum_{h\in\mathcal{H}} w_h(Y)\Big[(1-h_i)^2-(1-h_i)\Big]\mu_i\xi_i + \sigma^2\sum_{h\in\mathcal{H}} w_h(Y)h_i^2(1-\xi_i^2).$$

With this equation, using the same arguments as in proving (3.24), we arrive at

$$(1+2\gamma)\mathbf{E}^{1/m}\left[\sum_{h\in\mathcal{H}} w_h(Y)R(\mu,h)\right]^m \le \mathbf{E}^{1/m}\left[\sum_{h\in\mathcal{H}} w_h(Y)\|\mu - h\cdot Y\|^2\right]^m + \frac{Km\sigma^2}{\gamma}.$$

Therefore by Lemma 3.2 we have

$$\left\{\mathbf{E}_\mu^{1/m}\left[\sum_{h\in\mathcal{H}} w_h(Y)R(\mu,h)\right]^m\right\}^{1/2} \le \left\{\mathbf{E}_\mu^{1/m}\left[\sum_{h\in\mathcal{H}} w_h(Y)\|\mu - h\cdot Y\|^2\right]^m\right\}^{1/2} + \sigma\sqrt{Km}.$$

Next, substituting this equation in (3.24) and using that $(x+y)^2 \le 2x^2+2y^2$ we arrive at the following upper bound

$$\mathbf{E}_\mu^{1/m}[\delta_2(\mu)]_+^m \le 4\gamma\mathbf{E}_\mu^{1/m}\left[\sum_{h\in\mathcal{H}} w_h(Y)\|\mu - h\cdot Y\|^2\right]^m + \frac{Km\sigma^2}{\gamma}. \tag{3.25}$$

Let us now consider $\delta_3(\mu)$. We have by (3.8)

$$\begin{aligned}
&\log\left\{\sum_{h\in\mathcal{H}} \pi_h \exp\left[-\frac{\tilde{R}(Y,h)-\tilde{r}(Y,\mathcal{H})}{2\beta\sigma^2}\right]\right\} \\
&\ge \log\left\{\sum_{h\ge h^\circ(Y)} \pi_h \exp\left[-\frac{\tilde{R}(Y,h)-\tilde{r}(Y,\mathcal{H})}{2\beta\sigma^2}\right]\right\} \\
&= \log\left\{\sum_{h\ge h^\circ(Y)} \pi_h \exp\left[-\frac{\|(1-h)\cdot Y\|^2 - \|[1-h^\circ(Y)]\cdot Y\|^2}{2\beta\sigma^2}\right.\right. \\
&\quad \left.\left. -\frac{1}{\beta}\sum_{i=1}^n [h_i - h_i^\circ(Y)]\right]\right\} \\
&\ge \log\left\{\sum_{h\ge h^\circ(Y)} \pi_h \exp\left[-\frac{1}{\beta}\sum_{i=1}^n [h_i - h_i^\circ(Y)]\right]\right\} \ge 0
\end{aligned} \tag{3.26}$$

and hence

$$\delta_3(\mu) \leq 0. \tag{3.27}$$

The final step in the proof is to bound from above the last term at the right-hand side of Equation (3.19), namely $\delta_4(\mu)$. We will do this the help of Lemma 3.6. First of all note that for all $h > \hat{h}^\epsilon$, where $\hat{h}^\epsilon$ is defined by (3.3), we have

$$\tilde{R}(Y,h) - \tilde{r}(Y,\mathcal{H}) \geq 2\beta\epsilon\sigma^2\big[\|h\|^2 - \|h^\circ(Y)\|^2\big] + 2\beta\sigma^2.$$

In order to make use of Lemma 3.6, let us set

$$\begin{aligned} q_h =& \exp\left[-\frac{\tilde{R}(Y,h) - \tilde{r}(Y,\mathcal{H})}{2\beta\sigma^2}\right] = \exp\left[-\frac{\tilde{R}(Y,h) - \tilde{R}(Y,h^\circ)}{2\beta\sigma^2}\right], \\ \mathcal{G} =& \big\{h \in \mathcal{H} : \|h^\circ\|_1 \leq \|h\|_1 < \|h^\circ\|_1 + 1\big\}, \\ \widetilde{h} =& \hat{h}^\epsilon. \end{aligned}$$

By Condition (1.18) we have that for all $h \geq \hat{h}^\epsilon$

$$\begin{aligned} q_h \leq& \exp\left\{-\frac{2\sigma^2\beta K_\circ\epsilon(\|h\|_1 - \|h^\circ\|_1)}{2\beta\sigma^2} - 1\right\} \\ =& \exp\big\{-K_\circ\epsilon(\|h\|_1 - \|h^\circ\|_1) - 1\big\}. \end{aligned} \tag{3.28}$$

Note also that similar to (3.26) it can be checked easily that

$$q_h \geq \exp\left(-\frac{1}{\beta}\right), \ h \in \mathcal{G}$$

and thus it follows immediately from the definition of $\mathcal{G}$ and (3.11) that

$$\sum_{h\in\mathcal{G}} \pi_h q_h \geq \exp\left(-\frac{1}{\beta}\right) \sum_{h\in\mathcal{G}} \pi_h \geq \frac{1}{2\beta} \exp\left(-\frac{2}{\beta}\right)$$

and hence

$$E(\mathcal{G}, K_\circ\epsilon) \leq \frac{C(K_\circ,\beta)}{2\epsilon} + \frac{1}{2}\log\frac{C(K_\circ,\beta)}{\epsilon}. \tag{3.29}$$

Denote for brevity

$$\rho(\mu,\mathcal{H}) = \left[2 + \frac{r(\mu,\mathcal{H}}{\sigma^2}\right].$$

With Equations (3.29), (3.9), and Lemmas 3.6, 3.3 we obtain

$$\begin{aligned}
&\mathbf{E}_{\mu}^{1/m}\Big\{\delta_4(\mu)-\log[\rho(\mu,\mathcal{H})]\Big\}_{+}^{m}\\
&\leq 2\mathbf{E}_{\mu}^{1/m}\log^{m}\left\{\frac{\|\hat{h}^{\epsilon}\|+\exp[E(\mathcal{G},K_{\circ}\epsilon)/2]}{\sqrt{\beta K_{\circ}\rho(\mu,\mathcal{H})}}\right\}\\
&\leq 2\log\Bigg\{\frac{1}{\sqrt{1-2\beta\epsilon}}\sqrt{\frac{r(\mu,\mathcal{H})}{\sigma^2}}\\
&+\frac{K\sqrt{m+\beta}}{\sqrt{1-2\beta\epsilon}}+\sqrt{\frac{C(K_{\circ},\beta)}{\epsilon}}\exp\left[\frac{C(K_{\circ},\beta)}{2\epsilon}\right]\Bigg\}-\log\big[K_{\circ}\beta\rho(\mu,\mathcal{H})\big].
\end{aligned}\tag{3.30}$$

Our next step is to minimize the right-hand side at this equation in $\epsilon\in(0,1/(2\beta))$. Note that that for any $\epsilon\leq 1/(3\beta)$ we have

$$\begin{aligned}
\frac{1}{\sqrt{1-2\beta\epsilon}}\sqrt{\frac{r(\mu,\mathcal{H})}{\sigma^2}}+\frac{K\sqrt{m+\beta}}{\sqrt{1-2\beta\epsilon}}&\leq\sqrt{\frac{r(\mu,\mathcal{H})}{\sigma^2}}+K\sqrt{m+\beta}\\
&+K\epsilon\left[\sqrt{\frac{r(\mu,\mathcal{H})}{\sigma^2}}+K\sqrt{m+\beta}\right].
\end{aligned}\tag{3.31}$$

Consider the following function

$$\Psi(x)=\min_{\epsilon\in[0,1/(3\beta)]}\left\{\epsilon x+\sqrt{\frac{C(K_{\circ},\beta)}{\epsilon}}\exp\left[\frac{C(K_{\circ},\beta)}{2\epsilon}\right]\right\}.$$

It is clear that $\Psi(0)$ is bounded from above. It is also easy to check with $\epsilon=4C(K_{\circ},\beta)/\log(x)$ that for any $x\geq C(K_{\circ},\beta)$

$$\Psi(x)\leq\frac{4C(K_{\circ},\beta)x}{\log(x)}+\frac{\sqrt{x\log(x)}}{2}\leq\frac{C(K_{\circ},\beta)x}{\log(x)}.$$

Therefore, combining this equation with (3.30), (3.31), we arrive at

$$\mathbf{E}_{\mu}^{1/m}\Big\{\delta_4(\mu)-\log[\rho(\mu,\mathcal{H})]\Big\}_{+}^{m}\leq\log\big[C(K_{\circ},\beta)\sqrt{m+\beta}\big].\tag{3.32}$$

Next, substituting (3.23), (3.25), (3.27), and (3.32) in (3.19) we arrive at

$$\begin{aligned}
\mathbf{E}_{\mu}^{1/m}\left[\sum_{h\in\mathcal{H}}w_h(Y)\big\|\mu-h\cdot Y\big\|^2\right]^{m}&\leq r(\mu,\mathcal{H})+K\sigma\sqrt{mr(\mu,\mathcal{H})}\\
&+4\gamma\mathbf{E}_{\mu}^{1/m}\left[\sum_{h\in\mathcal{H}}w_h(Y)\|\mu-h\cdot Y\|^2\right]^{m}+\frac{Km\sigma^2}{\gamma}\\
&+2\beta\sigma^2\log\big[C(K_{\circ},\beta)(m+\beta)\big]+2\beta\sigma^2\log[\rho(\mu,\mathcal{H})].
\end{aligned}\tag{3.33}$$

Using this equation and minimizing the right-hand at (3.33) in $\gamma \in (0, 1/4)$, we obtain with the help of Lemma 3.2

$$\begin{aligned}\mathbf{E}_{\mu}^{1/(2m)}\Big[\sum_{h\in\mathcal{H}} w_h(Y)\|\mu - h\cdot Y\|^2\Big]^m &\le K\sqrt{m}\sigma \\ &+ \Big\{r^{\beta}(\mu,\mathcal{H}) + 2\beta\sigma^2\log\big[C(K_\circ,\beta)(m+\beta)\big]\Big\}^{1/2}.\end{aligned} \tag{3.34}$$

Similarly to (3.33), we have that for any $\gamma \in (0, 1/4)$

$$\begin{aligned}&\mathbf{E}_{\mu}^{1/m}\Big[\sum_{h\in\mathcal{H}} w_h(Y)\big\|\mu - h\cdot Y\big\|^2 - r^{\beta}(\mu,\mathcal{H})\Big]_+^m \\ &\quad\le K\sigma\sqrt{mr^{\beta}(\mu,\mathcal{H})} + 2\beta\sigma^2\log\big[C(K_\circ,\beta)(m+\beta)\big] \\ &\quad\quad + 4\gamma\mathbf{E}_{\mu}^{1/m}\Big[\sum_{h\in\mathcal{H}} w_h(Y)\|\mu - h\cdot Y\|^2\Big]^m + \frac{Km\sigma^2}{\gamma}.\end{aligned}$$

and substituting in this equation (3.34), we arrive at

$$\begin{aligned}&\mathbf{E}_{\mu}^{1/m}\Big[\sum_{h\in\mathcal{H}} w_h(Y)\big\|\mu - h\cdot Y\big\|^2 - r^{\beta}(\mu,\mathcal{H})\Big]_+^m \\ &\quad\le K\sigma\sqrt{mr^{\beta}(\mu,\mathcal{H})} + 2\beta\sigma^2\log\big[C(K_\circ,\beta)(m+\beta)\big] \\ &\quad\quad + \gamma\Big\{8r^{\beta}(\mu,\mathcal{H}) + 2\beta\sigma^2\log\big[C(K_\circ,\beta)(m+\beta)\big] + 16K\sigma^2 m\Big\} \\ &\quad\quad + \frac{Km\sigma^2}{\gamma}.\end{aligned} \tag{3.35}$$

Our final step is to minimize the right-hand side at this equation in $\gamma \in (0, 1/4)$. Notice that

$$\gamma_\circ = \sigma\sqrt{Km}\Big\{8r^{\beta}(\mu,\mathcal{H}) + 2\beta\sigma^2\log\big[C(K_\circ,\beta)(m+\beta)\big] + 16K\sigma^2 m\Big\}^{-1/2}$$

is the unconstrained minimizer of the right-hand side. However it is clear

that $\gamma_\circ \le 1/4$ and hence we have by (3.35)

$$
\begin{aligned}
&\mathbf{E}_\mu^{1/m}\biggl[\sum_{h\in\mathcal{H}} w_h(Y)\bigl\|\mu - h\cdot Y\bigr\|^2 - r^\beta(\mu,\mathcal{H})\biggr]_+^m \\
&\le K\sigma\sqrt{mr^\beta(\mu,\mathcal{H})} + 2\beta\sigma^2\log\bigl[C(K_\circ,\beta)(m+\beta)\bigr] \\
&\quad + K\sigma\sqrt{m}\Bigl\{r^\beta(\mu,\mathcal{H}) + 2\beta\sigma^2\log\bigl[C(K_\circ,\beta)(m+\beta)\bigr] + K\sigma^2 m\Bigr\}^{1/2} \\
&\le K\sigma\sqrt{m}\Bigl\{r^\beta(\mu,\mathcal{H}) + K\sigma^2 m\Bigr\}^{1/2} + K\beta\sigma^2\log\bigl[C(K_\circ,\beta)(m+\beta)\bigr] \\
&\le K\sigma\sqrt{mr^\beta(\mu,\mathcal{H})} + K\sigma^2 m + K\beta\sigma^2\log\bigl[C(K_\circ,\beta)\bigr] \\
&\quad + K\beta\sigma^2\log(m+\beta) \\
&\le K\sigma\sqrt{mr^\beta(\mu,\mathcal{H})} + K\sigma^2 m + K\beta[1+\log(\beta)]\sigma^2\log\bigl[C(K_\circ,\beta)\bigr].
\end{aligned}
\tag{3.36}
$$

In deriving this inequality it was used that

$$
2\epsilon\log(x) - x \le 2\epsilon[1+\log(\epsilon)],\ \ x>0,\ \ \epsilon>0. \tag{3.37}
$$

Finally, Equation (3.36) with (3.13), we finish the proof. ■

3.3 Proof of Theorem 1.4

It follows from (1.22) that for any $m \ge 1$

$$
\mathbf{E}^{1/m}\Bigl[\|\bar\mu^\beta(Y)-\mu\| - \sqrt{r^\beta(\mu,\mathcal{H})}\Bigr]_+^m \le K\sqrt{m}\sigma + \frac{K\sigma^2[C(K_\circ,\beta)+m]}{\sqrt{r^\beta(\mu,\mathcal{H})}}.
$$

Hence, by the Markov inequality we obtain

$$
\begin{aligned}
&\mathbf{P}_\mu\Bigl\{\|\bar\mu^\beta(Y)-\mu\| \ge \sqrt{r^\beta(\mu,\mathcal{H})} + x\Bigr\} \le \exp\Bigl\{-m\log(x) \\
&\quad + m\log\Bigl[\mathbf{E}^{1/m}\Bigl[\|\bar\mu^\beta(Y)-\mu\| - \sqrt{r^\beta(\mu,\mathcal{H})}\Bigr]_+^m\Bigr]\Bigr\} \\
&\le \exp\Bigl\{-m\log(x) + m\log\Bigl[K\sqrt{m}\sigma + \frac{K\sigma^2[C(K_\circ,\beta)+m]}{\sqrt{r^\beta(\mu,\mathcal{H})}}\Bigr]\Bigr\} \\
&\le \exp\Bigl\{-m\log\Bigl(\frac{x}{K\sigma}\Bigr) + \frac{m\log(m)}{2} + \log\Bigl[1 + \frac{K\sigma[C(K_\circ,\beta)+m]}{\sqrt{m}\sqrt{r^\beta(\mu,\mathcal{H})}}\Bigr]\Bigr\}.
\end{aligned}
$$

Therefore choosing

$$
m = \frac{x^2}{K^2\sigma^2\exp(1)},
$$

we get with (3.37) that for any $x \geq \sigma$

$$\begin{aligned}
&\mathbf{P}_\mu\left\{\|\bar{\mu}^\beta(Y)-\mu\| \geq \sqrt{r^\beta(\mu,\mathcal{H})}+x\right\} \\
&\quad\leq \exp\left\{-\frac{x^2}{K\sigma^2}+\log\left\{C(K_\circ,\beta)+\frac{K\sigma}{\sqrt{r^\beta(\mu,\mathcal{H})}}\left(\frac{x}{\sigma}\right)\right\}\right\} \\
&\quad\leq \exp\left\{-\frac{x^2}{K\sigma^2}+C(K_\circ,\beta)+\frac{1}{2}\log\left(\frac{x}{\sigma}\right)^2\right\} \\
&\quad\leq \exp\left\{-\frac{x^2}{K\sigma^2}+C(K_\circ,\beta)\right\}. \quad \blacksquare
\end{aligned}$$

References

[1] Alquier, P. and Lounici, K. (2011). Pac-bayesian theorems for sparse regression estimation with exponential weights. *Electronic Journal of Statistics* **5**, 127–145.

[2] Arias-Castro, E. and Lounici, K Variable Selection with Exponential Weights and ℓ_0-Penalization. arXiv:1208.2635

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle *Proc. 2nd Intern. Symp. Inf. Theory.* 267–281.

[4] Catoni, O. (2004). *Statistical learning theory and stochastic optimization.* Lectures Notes in Math. **1851** Springer-Verlag, Berlin.

[5] Chernousova, E., Golubev, Yu., and Krymova, E. (2013) Ordered Smoothers With Exponential Weighting. *Electronic J. Statist.* **7** 2395–2419.

[6] Dalayan, A. and Salmon, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.* **40** 2327–2355.

[7] Demmler, A. and Reinsch, C. (1975). Oscillation matrices with spline smoothing. *Numerische Mathematik.* **24** 375–382.

[8] Golubev, Yu. (2010). On universal oracle inequalities related to high dimensional linear models. *Ann. Statist.* **38** No. 5 2751–2780.

[9] Golubev, G. (2012). Exponential weighting and oracle inequalities for projection estimates. *Problems of Information Transmission*, No. 3, V. 48, 269–280.

[10] Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.* **28** 681–712.

[11] Kneip, A. (1994). Ordered linear smoothers. *Annals of Stat.* **22** 835–866.

[12] Lecué, G. (2007). Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.* **35** 1698–1721.

[13] Leung, G. and Barron, A. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory.* **52** 3396–3410.

[14] Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

[15] Nemirovski, A. (2000). *Topics in non-parametric statistics.* Lectures Notes in Math. **1738** Springer-Verlag, Berlin.

[16] Rigollet, P. and Tsybakov, A. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16** 260–280.

[17] Rigolet, P. and Tsybakov, A. (2012). Sparse estimation by exponential weighting. *Statistical Science*, Vol. 27, No. 4, 558–575.

[18] Stein, C. (1973). Estimation of the mean of a multivariate normal distribution. *Proc. Prague Symp. on Asymptotic Statistics*, 345–381, Prague, Czechoslovakia.

[19] Yang, Y. (2000). Combining different procedures for adaptive regression. *J. Multivariate Anal.* **74** 135–161.